基于双向 LSTM 和 GBDT 的中医文本关系抽取模型 *

罗计根，杜建强†，聂　斌，熊旺平，刘　蕾，贺　佳

(江西中医药大学 计算机学院, 南昌 330004)

摘　要：为解决采用 Softmax 作为长短期记忆网络分类器导致实体关系识别模型泛化能力不足，不能较好适用中医实体关系抽取等问题，提出一种融合梯度提升树的双向长短期记忆网络的关系识别算法(BILSTM-GBDT)。先采用 word2vec 对中医文本进行向量化表示，再利用基于注意力机制的双向长短期记忆网络提取高阶特征，最后采用集成分类模型梯度提升树作为特征分类器，提高关系识别效果。在中医等多个关系语料库上的实验结果表明，该模型与传统 SVM 方法、GBDT 方法及其深度学习方法相比，均有更高的精确率、召回率和 F 值。

关键词：关系抽取；LSTM；梯度提升树；注意力机制；中医文本

中图分类号：TP　　doi: 10.3969/j.issn.1001-3695.2018.07.0420

TCM text relationship extraction model based on bidirectional LSTM and GBDT

Luo Jigen[1], Du Jianqiang†, Nie Bin, Xiong Wangping, Liu Lei, He Jia

(*School of Computer Jiangxi University of Traditional Chinese Medicine*, *Nanchang* 330004, *China*)

Abstract: In order to solve the problem that the use of Softmax as a long-short-term memory network classifier leads to the lack of generalization ability of the entity relationship recognition model, it is not suitable for the extraction of TCM entity relationships. This paper proposed a bidirectional long short-term memory (BILSTM) relational identification algorithm (BILSTM-GBDT) that incorporates a gradient boosting decision tree (GBDT) . Firstly, The Chinese medicine text vector is trained by word2vec, then the high-order features are extracted by the Bidirectional Long Short-Term Memory network based on the attention mechanism. Finally, the integrated classification model gradient lifting tree is used as the feature classifier to improve the relationship recognition effect. Experimental results on multiple relational corpora such as Chinese medicine show that the model has higher accuracy, recall and F value than traditional SVM method, GBDT method and deep learning method.

Key words: relationship extraction; LSTM; GBDT; attention mechanism; Chinese medicine text

0　引言

中医诊断知识是中华民族千百年遗留下的瑰宝，对中医临床具有很强的指导作用。随着中医诊断数据的不断增加，语句表达形式上具有一定的灵活性，实体关系也随之变得复杂。中医实体关系识别[1,2]是中医领域信息抽取[3]的一部分，是指在给定实体对和非结构文本情况下，识别两者之间存在的语义关系。

例如下面这个句子包含方剂、中药、症状、舌像、脉象、证型等多类实体：补天大造丸由牛膝、当归、小茴香等多味中药组成，其主治为咳逆喘息少气，咯痰色白有沫，血色暗淡，潮热，自汗，盗汗，声嘶或失音，面浮肢肿，心慌，唇紫，肢冷，形寒，口舌生糜。苔黄而剥，舌质光淡，脉微细而数的肺痨阴阳两虚证。对于上面句子中，在关系抽取任务中，需要准确的识别出实体“补天大造丸”和实体“牛膝”、“当归”和“小茴香”之间的语义关系，其大类关系为“方药”，小类关系为“组成”，实体“补天大造丸”与实体“肺痨阴阳两虚证”是“治疗”关系，实体“肺痨阴阳两虚证”与实体“脉微细而数”是“脉象”，也就是脉微细而数是痨阴阳两虚证的脉象，此外还有证型和症状，证型和舌像的关系。整个句子的关系表现形式如图 1 所示。

1　相关工作

关系抽取对于信息检索、篇章理解、知识图谱构建等研究都具有及其重要的研究意义。目前较流行的关系抽取方法有：基于特征工程的抽取方法、基于核函数的抽取方法和基于深度

收稿日期：2018-07-31；**修回日期：**2018-09-13　**基金项目：**国家自然科学基金资助项目（61363042，61562045，61762051）；江西省科技厅重大研发计划资助项目（20171ACE50021）；江西省研究生创新专项资金资助项目（YC2017-S349）；江西省科技厅重点研发计划资助项目（20171BBG70108）

作者简介：罗计根（1991-），男，江西萍乡人，硕士研究生，主要研究方向为机器学习、自然语言处理；杜建强（1968-），男（通信作者），江西南昌人，教授，博士，主要研究方向为医药信息与数据挖掘（jianqiang_du@163.com）；聂斌（1972-）男，江西南昌人，副教授，硕士，主要研究方向为数据挖掘、机器学习；熊旺平（1982-），男，江西南昌人，副教授，硕士，主要研究方向为数据挖掘、机器学习；刘蕾（1991-），女，河北石家庄人，硕士研究生，主要研究方向为机器学习、自然语言处理；贺佳（1992-），女，陕西渭南人，硕士研究生，主要研究方向为机器学习、自然语言处理.

学习的抽取方法[4]。

在基于特征工程的抽取方法中，主要是利用词汇特征、句法特征和语义特征[5]。虽然基于特征工程的关系抽取方法在一定程度上取得了不错的效果，但是由于句子的表达形式越来越复杂，特征提取越来越困难，导致基于特征的关系抽取效果很难提升。基于核函数[6]的关系抽取方法不同于特征工程，它主要考虑的是句子本身的结构信息，不需要建立高维的数据特征向量。它使用句法结构树作为输入对象，通过核函数比较语料之间的结构相似性进行关系分类。但是由于句子隐形特征中存在人们无法识别的噪声，且语义相同存在不同的表达形式，句子的长短表达能力不一样导致基于核函数的关系抽取也存在一定的弊端。

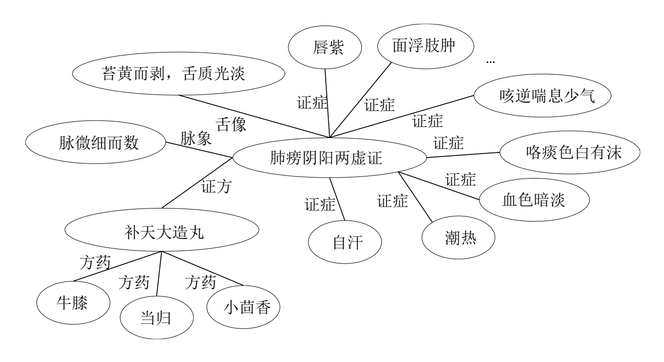


图 1　例句关系

随着深度学习的不断发展，以其自动提取特征的优势被更多的应用在关系抽取任务中[7,8]。Vu 等人[9]提出将深度循环神经网络（deep recurrent neural networks，DRNN）应用于关系抽取任务中，通过解析树的方式将句子分成两个部分，而后将其输入到多层循环神经网络中去。Zeng 等人[10]提出一种融合位置信息的卷积神经网络（convolutional neural networks，CNN）的关系抽取算法，为有效缓解长距离依赖问题，该算法考虑 N-gram 特征，但是由于 CNN 中的滤波器选择不能太大，导致不能完全长距离依赖的问题。LSTM 是由 Hochreiter 等人[11]提出的一种 RNN 改进模型，它设置了三种门限结构，通过记忆和遗忘等操作解决 RNN 和 CNN 存在的长距离依赖问题，目前也越来越多的应用在关系抽取任务上，Miwa 等人[12]利用 LSTM 引入 SPTree 中进行关系抽取。但是，以上模型都采用了 Softmax 作为分类器，导致实体关系识别模型泛化能力不足[13]，不能较好适应中医实体关系分类的问题。

为解决上述问题，本文提出一种融合梯度提升树(gradient boosting decision tree，GBDT)算法[14]的双向长短期记忆网络（bidirectional long short-term memory，BILSTM）模型。其中在使用 BILSTM 进行特征提取的同时，加入 Attention 机制抓取关键字词对句子理解[7]，解决该模型容易被无关词干扰的问题。特征提取后采用 GBDT 对关系分类训练预测，由于 GBDT 的基础模型具有低方差高偏差等优势，使得集成模型更具稳定性，可以在一定程度上解决采用 Softmax 作为长短期记忆网络分类器导致泛化能力不足的问题。

2　融合 GBDT 的 BILSTM 关系抽取

融合 GBDT 的 BILSTM 关系抽取模型，采用 BILSTM 模型获取前后两个方向的深层隐含特征，同时有效解决传统深度学习方法中长距离依赖的问题。同时在利用 BILSTM 提取特征时加入 Attention 机制考虑关键字词对特征的影响，从而获取更多的上下文信息；然后采用 GBDT 算法对提取的特征进行分类处理，得到最终的关系类别。该模型如图 2 所示，其主要包括两个部分：

a）加入 Attention 机制的 BILSTM 特征提取。将训练语料库的词向量输入到 BILSTM 模型中，采用 Attention 机制计算注意力概率，对 BILSTM 模型输入的关键字词重要性分析，根据注意力概率获取 BILSTM 模型的输出特征。

b）基于 GBDT 的关系分类。将 BILSTM 模型得到的特征输入到 GBDT 算法中，不断迭代构建决策树，利用上次模型的负梯度改进模型，在残差减少的梯度方向上建立新的决策树。

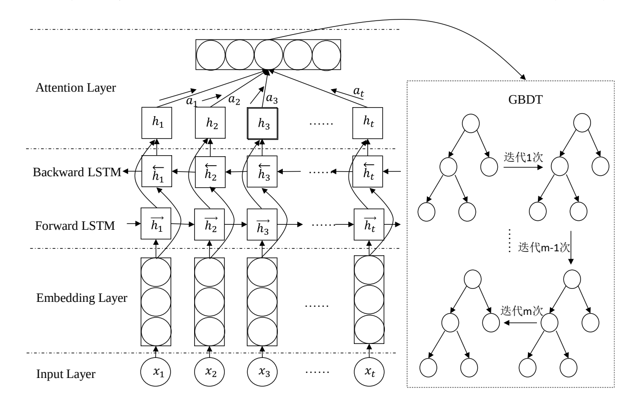


图 2　融合 GBDT 的 BILSTM 关系抽取模型

2.1　基于注意力机制的 BILSTM 特征提取

由于 LSTM 不能直接处理文本数据，需先利用 Google 的开源工具 Word2vec 将文本转换成字向量。假设输入的句子为 S，所包含的字集合为 $W(w_1,w_2,w_3,\dots,w_m)$，m 为句子长度，其中第 t 个字的字向量为：$w_t^* \in R^d$，上式中 d 为词向量的维度，则输入文本表示为

$$S=\left[w_1^*,w_2^*,\dots,w_m^*\right]\in R^{T\times d} \tag{1}$$

LSTM 神经网络是一种特殊的 RNN，其思想是用 LSTM 单元去替代 RNN 中隐含层的神经单元。LSTM 单元是由输入门（Input gate），输出门（Output gate）和遗忘门（Forget gate）三个门组成，由于 LSTM 的特殊结构，让 LSTM 神经网络可以在一定程度上解决长距离依赖问题。

在 t 时刻，LSTM 各单元组成部分的更新情况如下所示。

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{2}$$

$$i_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_i) \tag{3}$$

$$\tilde{c}_t=\tanh(W_c\cdot[h_{t-1},x_t]+b_c) \tag{4}$$

$$c_t=f_t\circ c_{t-1}+i_t\circ\tilde{c}_t \tag{5}$$

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o) \tag{6}$$

$$h_t=o_t\circ\tanh(c_t) \tag{7}$$

其中：σ 表示 sigmoid 激活函数，$\circ$ 是元素乘，x_t 为 t 时刻 LSTM

的输入向量，h_t 代表了隐含状态，W_f，W_i，W_c，W_o 分别代表了遗忘门，输入门，记忆单元，输出门的权值矩阵。b_f，b_i，b_c，b_o 分别代表了遗忘门，输入门，记忆单元，输出门的偏置。f_t，i_t，c_t，o_t 表示遗忘门，输入门，记忆单元状态和输出门。

为充分利用上下文信息，挖掘更多的隐含特征，有效解决关系抽取问题，本文设计双向 LSTM 神经网络，该方法是由两个相反方向的 LSTM 神经网络组成，其模型结构如图 1 中 BILSTM layer 部分所示，其中 $\vec{h}_t$ 是前向 LSTM 神经网络在 t 时刻的输出，$\overleftarrow{h}_t$ 是后向 LSTM 单元在时刻 t 的输出，所以时刻 t 的输出为前向后向的拼接，即 $h_t=[\vec{h}_t,\overleftarrow{h}_t]$。

由于每个字词对句子所属类别的贡献能力不同，利用 Attention 机制[15]的思想对句子进行更深的特征提取，提高关系分类精确率。例如在句子"现代研究：许氏报告用通脉四逆汤加味治疗少阴格阳证 16 例，全部治愈"中，普通的 BILSTM 神经网络对句子中的每一个词都是同等对待的，引入 Attention 机制后，模型通过注意力权值分配，重点关注"治疗"这个关键词。图 2 中 Attention layer 所示为在 BISLTM 模型后接入 Attention 机制的结构示意图。经过 Attention 层得到的全局输出向量为 H，则相关计算如下：

$$u_t=\tanh(w_w h_t+b_w) \tag{8}$$

$$a_t=softmax(u_t^T,u_w) \tag{9}$$

$$H=\sum_t a_t h_t \tag{10}$$

其中：u_t 是 h_t 的隐藏单元，u_w 为句子的上下文向量，a_t 为注意力向量，h_t 为 BILSTM 的输出向量，也即 $h_t=[\vec{h}_t,\overleftarrow{h}_t]$，$w_w$，$b_w$ 为注意力权重值和偏置，随机初始化并在训练中不断学习。

2.2　基于 GBDT 的关系分类

关系抽取可以看成是多分类问题，蓂雨洁等人[16]提出将 GBDT 用于微博立场检测当中，通过对语料库手动提取特征，完成文本分类。段大高等人[17]提出一种基于 GBDT 的虚假消息检测方法，通过提出评论中文本内容、用户属性，信息传播和时间等特征，利用 GBDT 实现分类。GBDT 是一种集成学习器，采用 Boosting 的思想，构造 N 个弱学习器，经过多次迭代形成最终强学习器。它采用的弱学习器为 CART 回归树，每一次迭代都是为了减少上一个模型的残差，并在残差减少的梯度上训练建立新的模型。

由 Attention 机制的 BILSTM 模型得到的特征和原始类别标签形成 GBDT 训练集数据 $T=\{(x_1,y_1),(x_2,y_2),(x_3,y_3),...,(x_n,y_n)\}$，其中 x_i 为语料库中第 i 个句子在模型提取的特征向量。y_i 的取值为关系类别。假设 GBDT 损失函数为 $L(y,f)$，则其表达式为

$$L(y,f)=\sum_{i=1}^{m}L(y_i,f(x_i)) \tag{11}$$

第 t 轮的第 i 个样本的损失函数的负梯度表示为

$$r_{ti}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)} \tag{12}$$

利用 (x_i,r_{tj}) 可以拟合一颗 CART 回归树，得到了第 t 棵决策树，其对应的叶节点区域。其中 J 为叶子节点的个数。

针对每一个叶子节点里的样本，求出使损失函数最小，也就是拟合叶子节点最好的输出值 c_{tj} 如下：

$$c_{tj}=\underbrace{\arg\min}_{c}\sum_{x_i\in R_j}L(y_k,f_{t-1}(x_i)+c) \tag{13}$$

得到本轮的决策树拟合函数为：

$$h_t(x)=\sum^{J}c_{tj}I(x\in R_{tj}) \tag{14}$$

本轮的最终强学习模型的表达式如下：

$$f_t(x)=f_{t-1}(x)+\sum^{J}c_{tj}I(x\in R_{tj}) \tag{15}$$

2.3　融合 GBDT 的 BILSTM 关系抽取模型

在实现实体关系抽取时，利用基于 Attention 机制的 BILSTM 抽取文本特征向量，得到特征组合 v，之后采用梯度提升树对特征组合进行分类训练和预测，得到最终每个句子的关系类别。BILSTM-GBDT 的优点在于可以在一定程度上解决传统深度学习方法在处理关系抽取时出现的泛化能力不强的问题，同时提高关系抽取的精确率。

BILSTM-GBDT 的具体算法流程如下：

a)利用 word2vec 对训练集样本进行 Embedding 操作，则每个输入句子的向量矩阵为：$S=\left[w_1^*,w_2^*,w_3^*,...,w_m^*\right]$；

b)将 $S=\left[w_1^*,w_2^*,w_3^*,...,w_m^*\right]$ 矩阵输入到 BILSTM 模型中，计算 t 时刻的正向输出 $\vec{h}_t$，逆向输出为 $\overleftarrow{h}_t$，BILSTM 层的输出特征为 $h_t=[\vec{h}_t,\overleftarrow{h}_t]$；

c)初始化 Attention 层中各节点的注意力权值，通过式（9）得到注意力概率 a_t，由式（10）计算得到最终输出特征 H；

d)利用 Attention 层输出 $H=\{h_1,h_2,h_3,...,h_n\}$ 和类别标签 $Y=\{y_1,y_2,y_3,...,y_m\}$ 构建梯度提升树。此时假设 GBDT 损失函数为 $L(y,f)=\sum_{i=1}^{m}L(y_i,f(x_i))$；

e)第 t 轮的第 i 个样本的损失函数的负梯度表示为：$_{tj}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)}$

f)拟合叶子节点输出值 c_{tj} 如下：$c_{tj}=argmin_{c}\sum_{x_i\in R_j}L(y_k,f_{t-1}(x_i)+c)$

g)决策树拟合函数为 $h_t(x)=\sum_{j=1}^{J}c_{tj}I(x\in R_{tj})$，经过 m 次迭代形成的集成学习模型为 $f_t(x)=f_{t-1}(x)+\sum_{j=1}^{J}c_{tj}I(x\in R_{tj})$。

h)通过模型迭代训练得到最终关系类别。

3　实验

3.1　实验设置

为验证本文提出的BILSTM-GBDT实体关系抽取算法的有效性，使用整理后的中医关系语料库（TCM RECoupus）对模型进行验证。中医关系语料库的来源包含了中医古籍文本、教学书籍、中医科研论文，在构建中医关系语料库时，先从标注文档中抽取出所需实体对，然后对文档进行断句处理，最后形成的语料库一共为 26855 个句子，包含 11 个类别。另外为验证本文提出改进算法的性能，本文除了使用中医关系语料库，还在 SemEval-2010 和 ACL2007 关系语料库上进行了对比实验。三组语料库详细情况如表 1 所示。

表 1　三种语料库信息

TCM RECoupus		SemEval-2010		ACL2007	
关系	数量	关系	数量	关系	数量
方药	3850	其他	1864	制造使用	1440
药症	2982	因果	1331	类属	1776
药证	1839	整体与部分	1253	转喻	470
方证	3642	实体与目标	1137	组织从属	2460
方症	2869	实体与来源	974	局部整体	981
方病	1683	生产者产品	948	人物	3116
病症	2258	会员与组织	923	地理位置	2157
证症	2754	实体与主题	895	-	-
药症	2115	内容与包含	732	-	-
舌证	2863	工具使用者	660	-	-
脉证	2651	-	-	-	-

本文通过将语料库按照每种关系的 7:3 方式划分模型的训练集和测试集。在 BILSTM 参数设置见表 2 所示，其中 Dropout、学习率和优化器等参数通过多组实验对比得出。

表 2　LSTM 神经网络参数设置

超参数	调参值
learning rate	0.001
dropout	0.5
gradient clipping	5.0
embedding-dim	300
optimizer	Adam
batch-size	64
hidden-dim	300
epoch	30

为体现本文提出的模型优势，采用准确率（P）、召回率（R）、F 值作为模型评价准则。

3.2　实验结果

利用中医实体关系语料库的训练集数据进行模型训练，用测试集数据进行 BILSTM-GBDT 模型评测，得到 11 类关系的精确率、召回率和 *F* 值如表 3 所示。

表 3　BILSTM-GBDT 模型实验结果

关系类别	*P*(%)	*R*(%)	*F*(%)
方药	93.25	92.53	92.88
药症	86.53	85.21	85.86
药证	85.12	86.74	85.92
方证	92.41	93.85	93.12
方症	83.96	81.23	82.57
方病	81.29	82.56	81.92
病症	88.20	87.06	87.62
证症	91.22	89.38	90.29
药症	84.23	82.71	83.46
舌证	90.36	89.27	89.81
脉证	91.28	92.06	91.67
所有	**87.98**	**87.51**	**87.74**

由表 3 可知，对于自定义的 11 种中医关系，其中方药、方证、证症、脉证四类关系的精确率、召回率和 F 值达到 90%以上。这是因为中医在这四类关系上表达，形式简单，实体形式固定，语料占比相对其他关系较大，所以效果明显优于其他类关系的实验效果。

为验证本文提出改进算法的性能，同时引入四种目前流行的关系抽取算法，分别为支持向量机（SVM）、梯度提升树（GBDT）、深度学习方法 BILSTM、融入注意力机制的 BILSTM 模型（BILSTM-ATT），这两类深度学习关系抽取模型都是在提取特征后采用 Softmax 进行关系分类。对比实验结果如表 4 和图 3 所示。

表 4 五种关系模型实验对比

	TCM RECoupus (%)			SemEval-2010			ACL2007		
方法	P	R	F	P	R	F	P	R	F
GBDT	79.21	78.63	78.92	79.53	81.21	80.36	78.33	79.61	78.96
SVM	75.63	78.57	77.07	78.92	79.34	79.13	77.63	76.97	77.29
BILSTM	83.78	82.35	83.06	81.66	82.62	82.14	82.35	82.74	82.54
BILSTM-ATT	85.82	85.44	85.63	84.09	83.25	83.67	84.67	84.14	84.40
BILSTM-GBDT	87.98	**87.51**	**87.74**	**86.13**	**85.52**	**85.82**	**85.85**	**86.28**	**86.06**

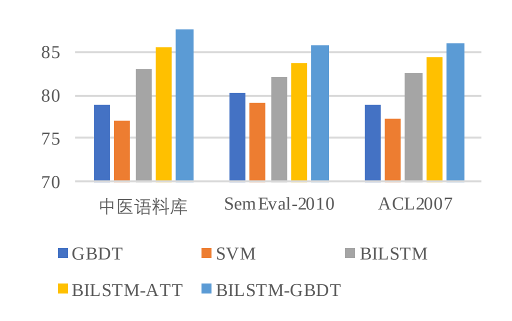


图 3　各组数据实验结果 F 值对比

结合表 4 和图 3 的实验结果，在中医关系语料库上集成算法 GBDT 在精确率、召回率和 *F* 值上都超过 SVM，可见集成学习算法相对传统机器学习算法增强了模型抗干扰能力，增强

了模型的泛化能力。但综合来看，BILSTM 比 GBDT 算法在 *F* 值高出 4.14%，说明 BILSTM 自动提取的句子深层特征有利用关系分类，可以得到更好的实验结果。BILSTM-ATT 在 BILSTM 的基础上增加了注意力机制，由实验结果可以看出，BILSTM-ATT 模型的 *F* 值相比 BILSTM 高出了 2.17%，正是加入了注意力机制的原因，它为每个输入的字向量提供权值，句子最终的特征通过此权重加权之后的整合，可以减少句子中噪声词的影响。BILSTM-GBDT 是在引入注意力机制的情况下，采用 GBDT 作为模型的分类器，在 F 值上相对 BILSTM-ATT 提高了 2.11%。

五种算法在 SemEval-2010 语料库上的实验结果见表 4 和图 3。由实验结果数据可知，集成算法 GBDT 在精确率、召回率和 *F* 值上依旧超过了 SVM，可见 GBDT 集成算法在关系分类上的优势。BILSTM-GBDT 在三个评价指标相对其他算法都最高的，在 *F* 值测评上高出 BILSTM-ATT 模型 2.25%。

根据表 4 和图 3 中五种算法在 ACL2007 语料库上的实验结果可知，集成算法 GBDT 相对 SVM 来说，在三个评测标准上仍有很大的优势。BILSTM-GBDT 相对其他两种以 Softmax 作为分类器的模型来说具有很大的优势，在 F 值评测上，BILSTM-GBDT 达到 86.06%，高出 BILSTM-ATT 模型 1.66%。为探究梯度提升树构建棵数 *m* 值和模型效果的问题，本文在三个关系语料库上做了相关实验，实验结果如图 4 所示，图中决策树的数目呈 10 的倍数递增，共计 10 次。

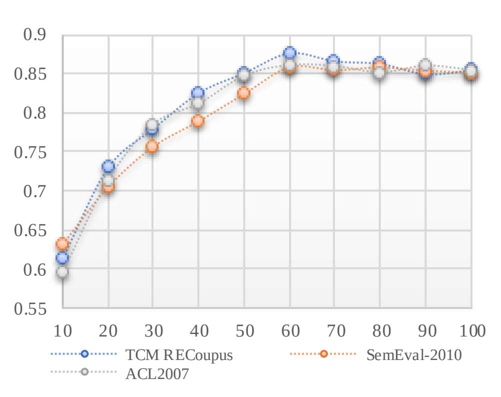


图 4 模型效果随 *m* 值变化

由图 4 可知，初始时随着决策树数目 *m* 的值增加，BILSTM-GBDT 模型在三个语料库上的 F 值均呈现上升的趋势，当梯度提升树构建到 60 棵的时候，此时模型效果达到最优，之后三个语料库上的 F 值基本趋于稳定，甚至会出现一定程度上的下降。

综上所述，BILSTM-GBDT 利用融合注意力机制的双向长短期记忆网络充分提取句子特征，使用集成学习 GBDT 作为分类器，在一定程度上解决了传统以 Softmax 用作分类器带来的模型泛化能力不强的问题，使实验结果更加稳定。相比来说，SVM 算法实验效果最差，使用 GBDT 这种集成学习算法使得模型更加稳定，泛化能力得到加强。但是由于人工提取的特征使 GBDT 跟 BILSTM-GBDT 的实验结果还是有一定差距。BILSTM-GBDT 先利用融合注意力机制的 BILSTM 模型提取高阶特征，然后利用集成学习 GBDT 迭代形成多棵决策树，增强了模型泛化能力，当梯度提升树构建到 60 棵的时候模型的效果趋于稳定。在三个语料库上的实验效果表明，BILSTM-GBDT 较 BILSTM-ATT 有明显的优势，在三个语料库上的综合表现 *F* 值分别提高了 2.11%、2.25%、1.66%，实验结果验证了该方法在解决以 Softmax 作为分类器带来的泛化能力不强问题上的有效性。

4 结束语

本文针对关系抽取任务上采用 Softmax 作为长短期记忆网络分类器导致模型泛化能力不足，不能较好适应中医实体关系分类的问题，提出了一种融合梯度提升树算法的双向长短期记忆模型，充分在利用双向 LSTM 自动提取特征的优势，并结合 Attention 机制抓取关键字词对句子理解，解决模型容易被无关词干扰的问题，最后利用 GBDT 低方差高偏差的优势，增强模型的鲁棒性和泛化性。通过对中医关系语料库和其他两个公开领域语料库实验的比较，证明本文提出的改进模型在准确率、召回率和 F 值上均有明显提高，是一种适合于中医特定领域的关系抽取模型。但是改进算法仍有不足之处，只能抽取预先定义好的关系，在接下来的工作中，将进一步研究对于新关系的抽取，以及如何将其扩展到其他领域中。

参考文献：

[1] 彤博辉，付琨，黄宇，等. 基于多通道卷积神经网的实体关系抽取 [J]. 计算机应用研究, 2017, 34 (3): 689-692. (Rong Bohui, Fu Kun, Huang Yu, *et al.* Relation extraction based on multi-channel convolutional neural network [J]. Application Research of Computers, 2017, 34 (3): 689-692.

[2] 陈宇，郑德权，赵铁军. 基于 deep belief nets 的中文名实体关系抽取 [J]. 软件学报, 2012, 23 (10): 2572-2585. (Chen Yu, Zheng Dequan, Zhao Tiejun. Chinese relation extraction based on deep belief nets [J]. Journal of Software, 2012, 23 (10): 2572-2585)

[3] 郭喜跃，何婷婷. 信息抽取研究综述 [J]. 计算机科学, 2015, 42 (2): 14-17. (Guo Xiyue, He Tingting. Survey about research on information extraction [J]. Computer Science, 2015, 42 (2): 14-17.)

[4] 段利国，徐庆，李爱萍，等. 实体词语义信息对中文实体关系抽取的作用研究 [J]. 计算机应用研究, 2017, 34 (1): 141-146. (Duan Liguoit, Xu Qing, Li Aiping, *et al.* Research on effect of entities semantic information on Chinese entity relation extraction [J]. Application Research of Computers, 2017, 34 (01): 141-146.)

[5] 甘丽新，万常选，刘德喜，等. 基于句法语义特征的中文实体关系抽取 [J]. 计算机研究与发展, 2016, 53 (2): 284-302. (Gan Lixin, Wan Changxuan, Liu Dexi, *et al.* Chinese named entity relation extraction based on syntactic and semantic features [J]. Journal of Computer Research & Development, 2016, 53 (2): 284-302.)

[6] 陈鹏，郭剑毅，余正涛，等. 基于凸组合核函数的中文领域实体关系抽取 [J]. 中文信息学报, 2013, 27 (5): 144-148. (Chen Peng, Guo Jianyi, Yu Zhengtao, *et al.* Chinese field entity relation extraction based on convex combination kernel function [J]. Journal of Chinese Information Processing,

2013, 27 (5): 144-148)

[7] Wang Yequan, Huang Minlie, Zhu Xiaoyan, *et al.* Attention-based LSTM for Aspect-level Sentiment Classification [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2017: 606-615.

[8] Chen Qian, Zhu Xiaodan, Ling Zhenhua, *et al.* Enhanced LSTM for natural language inference [C]// Proc of Meeting of the Association for Computational Linguistics. 2017: 1657-1668.

[9] Vu N T, Adel H, Gupta P, *et al.* Combining recurrent and convolutional neural networks for relation classification [EB/OL]. (2016-05-24) . https://arxiv. org/abs/1605. 07333.

[10] Zeng Daojian, Liu Kang, Lai Siwei, *et al.* Relation classification via convolutional deep neural network [C]// Proc of the 25th International Conference on Computational Linguistics. 2014: 2335-2344

[11] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9 (8): 1735-1780.

[12] Miwa Makoto, Bansal Mohit. End-to-end relation extraction using LSTMs on sequences and tree structures [EB/OL]. (2016-06-08) . https://arxiv. org/abs/1601. 00770.

[13] 胡杰，李少波，于丽娅，等. 基于卷积神经网络与随机森林算法的专利文本分类模型 [J]. 科学技术与工程, 2018, 18 (6): 268-272. (Hu Jie, Li Shaobo, Yu Liya, *et al.* A patent classification model based on convolutional neural networks and rand forest [J]. Science Technology and Engineering, 2018, 18 (6): 268-272)

[14] Jerome H. Friedman. greedy function approximation: a gradient boosting machine [J]. Annals of Statistics, 2001, 29 (5): 1189-1232.

[15] Baziotis C, Pelekis N, Doulkeridis C. DataStories at SemEval-2017 task 4: deep LSTM with attention for message-level and topic-based sentiment analysis [C]// Proc of International Workshop on Semantic Evaluation. 2017: 747-754.

[16] 奠雨洁，金琴，吴慧敏. 基于多文本特征融合的中文微博的立场检测 [J]. 计算机工程与应用, 2017, 53 (21): 77-84. (Dian Yujie; Jin Qin; Wu Huimin. Stance detection in Chinese microblogs via fusing multiple text features [J]. Computer Engineering and Applications, 2017, 53 (21): 77-84.)

[17] 段大高，盖新新，韩忠明，等. 基于梯度提升决策树的微博虚假消息检测 [J]. 计算机应用, 2018, 38 (2): 410-414. (Duan Dagao, Gai Xinxin, Han Zhongming, Liu Bingxin. Micro-blog misinformation detection based on gradient boost decision tree [J]. Journal of Computer Applications, 2018, 38 (2): 410-414.)